DOKUMEN SPESIFIKASI PROYEK *BIG DATA*
Optimasi Prediksi Produksi Padi & Beras Sumatera via Big Data Pipeline (Random Forest & Power BI)

Kelompok 19

Pardi Octaviando	122450132
Kholisaturrohmah	120450019
Elilya Octaviani	122450009
Tria Yunanni	122450062
Rut Junita Sari Siburian	122450103

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2025

1. Latar Belakang

Indonesia, sebagai negara dengan potensi agraris yang melimpah dan wilayah daratan yang luas, memberikan kesempatan besar bagi penduduknya untuk menggantungkan mata pencaharian pada sektor pertanian. Di antara gugusan pulau yang ada, Pulau Sumatera memegang peranan strategis sebagai salah satu lumbung padi dan beras nasional, dengan kontribusi signifikan terhadap pemenuhan kebutuhan pangan dalam negeri [1]. Sektor pertanian di Indonesia tidak hanya menjadi tulang punggung kehidupan masyarakat, tetapi juga memiliki daya ungkit yang besar dalam memperkuat stabilitas dan akselerasi pertumbuhan ekonomi makro [2]. Padi dan beras, selain berstatus sebagai komoditas pangan pokok yang kandungan karbohidratnya esensial sebagai sumber energi [3], juga merupakan tumpuan pendapatan utama bagi jutaan keluarga petani di Indonesia.

Meskipun Pulau Sumatera memiliki potensi besar, tantangan seperti variabilitas hasil produksi padi dan beras dari tahun ke tahun di berbagai provinsinya masih menjadi isu krusial. Fluktuasi ini seringkali mencerminkan adanya keterbatasan dalam pengelolaan data pertanian secara komprehensif, perencanaan tanam yang kurang optimal, hingga strategi distribusi pangan yang belum sepenuhnya adaptif. Untuk mengatasi kompleksitas ini, diperlukan sebuah sistem peramalan produksi padi dan beras yang andal dan terintegrasi. Sistem semacam ini krusial untuk mendukung pemerintah dan pemangku kepentingan lainnya dalam pengambilan keputusan yang lebih tepat sasaran terkait kebijakan pasokan pangan, optimalisasi jalur distribusi, serta upaya menjaga stabilitas harga di tingkat pasar. Berbagai jenis peramalan, baik produksi maupun harga komoditas pertanian seperti beras, menjadi informasi penting bagi pemerintah dalam merumuskan kebijakan [4]. Oleh karena itu, sistem peramalan produksi yang akurat, seperti yang diusulkan dalam penelitian ini, akan melengkapi upaya tersebut.

Tantangan dalam pengelolaan pertanian modern, seperti dampak perubahan iklim, aksesibilitas informasi yang belum merata bagi petani, serta tuntutan peningkatan produktivitas yang berkelanjutan, semakin mempertegas urgensi adopsi teknologi. Menjawab kebutuhan tersebut, penelitian ini mengusulkan pendekatan **"Optimasi Prediksi Produksi Padi & Beras Sumatera via Big Data Pipeline (Random Forest & Power BI)"**. Solusi komprehensif ini dirancang untuk meningkatkan akurasi peramalan dengan memanfaatkan data historis dan potensi data *real-time* melalui alur kerja yang sistematis dan terintegrasi. Agar proses peramalan dapat berjalan optimal, terutama saat menangani data pertanian dalam

volume besar, diperlukan sistem manajemen data yang tangguh. Penggunaan teknologi Big Data seperti Apache Hadoop dan Hive telah terbukti efektif dalam analisis data skala besar dan mendukung berbagai metode peramalan, tidak hanya terbatas pada sektor pertanian tetapi juga pada sektor lain seperti UMKM [5].

Pendekatan yang diusulkan dalam penelitian ini akan melibatkan perancangan dan implementasi **arsitektur *data warehouse*** dengan Skema Bintang menggunakan Apache Hive, yang memungkinkan penyimpanan dan analisis data produksi pertanian secara terstruktur dan historis dari seluruh Pulau Sumatera untuk periode 2018-2024. Kemampuan HiveQL dalam lingkungan Hadoop untuk melakukan *Exploratory Data Analysis* (EDA) pada dataset kompleks telah ditunjukkan dalam berbagai studi kasus, memberikan landasan untuk penggalian wawasan awal dari data pertanian ini [6]. Selanjutnya, akan dibangun sebuah ***big data pipeline*** yang efisien menggunakan Apache Hadoop untuk aspek penyimpanan data terdistribusi (HDFS) dan Apache Spark untuk keseluruhan proses ETL (*Extract, Transform, Load*), termasuk transformasi data mentah luas panen dan produksi padi menjadi format yang siap untuk dianalisis lebih lanjut. Inti dari kemampuan prediktif sistem ini terletak pada pengembangan **model menggunakan algoritma *Machine Learning* Random Forest Regression** melalui pustaka Spark MLlib, yang bertujuan menghasilkan estimasi produksi padi dan beras dengan tingkat akurasi yang lebih tinggi. Sebagai pelengkap, hasil analisis tren, performa model, dan visualisasi prediksi akan disajikan melalui **dashboard interaktif yang dikembangkan menggunakan Microsoft Power BI**, sehingga memudahkan interpretasi dan pemanfaatan informasi secara efektif oleh para pengambil kebijakan dan pelaku agribisnis.

Dengan mengintegrasikan kapabilitas Apache Hadoop untuk manajemen data skala besar, Apache Spark untuk pemrosesan data yang cepat dan analitik canggih, serta Apache Hive untuk fungsionalitas *data warehousing*, diharapkan sistem peramalan yang dihasilkan melalui *big data pipeline* ini mampu memberikan kontribusi signifikan terhadap perencanaan pertanian yang lebih efektif dan pengambilan kebijakan yang berbasis data kuat di tingkat regional Sumatera, demi mendukung ketahanan pangan dan kesejahteraan petani.

2. Tujuan

Tujuan dilakukannya penelitian ini diantaranya yaitu:

1. +Merancang **arsitektur *data warehouse*** yang optimal untuk pengelolaan data historis produksi padi dan beras di Pulau Sumatera (periode 2018-2024).
2. Mengimplementasikan ***big data pipeline*** yang efisien menggunakan **Hadoop (HDFS), Apache Hive (HiveQL), dan Apache Spark** untuk proses *Extract, Transform, Load* (ETL) dan pemrosesan data produksi.
3. Membangun **model prediksi produksi padi dan beras** menggunakan algoritma *Machine Learning* **Random Forest Regression** dengan memanfaatkan pustaka Spark MLlib.
4. Melakukan **evaluasi performa model prediksi** *Random Forest Regression* menggunakan metrik **Mean Absolute Percentage Error (MAPE)** dan **Root Mean Squared Error (RMSE)** untuk mengukur akurasi dan keandalannya.
5. Menyajikan hasil analisis tren produksi dan visualisasi hasil prediksi model secara interaktif melalui **dashboard Power BI** untuk mendukung pengambilan keputusan strategis.

3. Metodologi Penelitian

Rangkaian kerja penelitian ini disusun secara sistematis untuk mencapai sasaran utama, yakni optimalisasi peramalan kuantitas produksi padi dan beras di wilayah Pulau Sumatera. Keseluruhan proses operasional akan dieksekusi dalam suatu lingkungan *Hadoop Stack* yang ter-kontainerisasi menggunakan Docker, dirancang khusus untuk menangani analisis data secara *batch*. Berikut adalah tahapan metodologis yang akan ditempuh:

3.1. Perancangan Arsitektur Sistem (Gudang Data)

Projek ini mengadopsi pendekatan arsitektur **Medallion Pipeline** untuk pengelolaan data produksi komoditas padi dan beras secara bertahap, mulai dari data mentah hingga data yang siap untuk analisis dan pemodelan. Arsitektur ini memastikan kualitas, reliabilitas, dan tata kelola data yang baik melalui tiga lapisan utama: Bronze, Silver, dan Gold. Lapisan akhir (Gold) akan menyediakan data terstruktur yang dapat diorganisir menggunakan prinsip-prinsip gudang data, seperti Skema Bintang, untuk analisis tren, pelaporan, dan input bagi model machine learning.

Konfigurasi Lapisan dalam Medallion Pipeline:

1. **Lapisan Bronze (Raw Data Storage):**

a. **Tujuan:** Menyimpan data mentah (raw data) dari sumbernya dengan transformasi minimal. Lapisan ini berfungsi sebagai *data lake* awal atau *landing zone* untuk data historis dan potensi data *real-time* di masa depan.
b. **Sumber Data:** Data primer produksi padi dan beras dari BPS (dalam format CSV) untuk periode 2018-2024.
c. **Proses:** Data diekstrak dari file CSV dan dimuat ke dalam Hadoop Distributed File System (HDFS). Pada lapisan ini, data disimpan dalam format aslinya atau dengan konversi format file dasar (misalnya, ke Parquet untuk efisiensi penyimpanan dan kompresi) tanpa perubahan struktur atau konten signifikan, mempertahankan jejak audit data sumber.
d. **Penyimpanan:** HDFS.
e. **Tools:** Apache Hadoop (HDFS).

2. **Lapisan Silver (Cleansed, Conformed, and Enriched Data):**

a. **Tujuan:** Menyediakan data yang telah dibersihkan, divalidasi, distandarisasi, dikonfirmasi, dan diperkaya (*enriched*). Data pada lapisan ini lebih terstruktur, memiliki kualitas yang lebih baik, dan siap untuk berbagai keperluan analitik dasar serta menjadi sumber untuk lapisan Gold.
b. **Proses (Transformasi via Apache Spark):**
    1. Data dari Lapisan Bronze akan ditarik dan diproses menggunakan Apache Spark.
    2. **Pembersihan Data:** Penanganan nilai yang hilang (*missing values*) dengan strategi imputasi yang sesuai, identifikasi dan penanganan data yang menyimpang (anomali atau pencilan), serta eliminasi rekaman data yang redundan.
    3. **Transformasi Struktur:** Restrukturisasi data (misalnya, *unpivoting* kolom-kolom data tahunan menjadi format baris individual), standarisasi tipe data, dan penamaan kolom yang konsisten.
    4. **Validasi dan Konformasi:** Memastikan integritas data, misalnya kesesuaian format tanggal, validasi kode wilayah.
    5. **Enrichment:** Penggabungan dengan data referensi lain jika ada, atau penambahan atribut turunan sederhana.
    6. **Pembentukan Tabel Terstruktur:** Data yang telah ditransformasi akan dimuat ke dalam tabel-tabel Apache Hive. Pada tahap ini, struktur yang mengarah ke model analitik seperti Skema Bintang mulai dibentuk untuk mendukung kueri analitik.
c. **Struktur Data (di Apache Hive):** Meskipun belum sepenuhnya menjadi Skema Bintang yang matang untuk BI, tabel-tabel di lapisan Silver akan lebih terorganisir dan mungkin mencakup versi awal dari:
    1. Tabel Dimensi (misalnya, Dim_Waktu_Silver, Dim_Lokasi_Silver, Dim_Komoditas_Silver): Menyajikan konteks temporal, spasial, dan komoditas yang telah dibersihkan.
    2. Tabel Fakta Parsial/Dasar (misalnya, Fact_Produksi_Silver): Berisi metrik produksi yang telah dibersihkan seperti Luas_Panen_Aktual

dan Jumlah_Produksi_Aktual, terhubung dengan tabel dimensi versi Silver.

d. **Penyimpanan:** HDFS (data untuk tabel Hive).
e. **Tools:** Apache Spark (untuk ETL), Apache Hive (untuk definisi skema dan penyimpanan data terstruktur).

3. **Lapisan Gold (Aggregated and Application-Specific Data):**

a. **Tujuan:** Menyediakan data yang sangat terkurasi, teragregasi, dan dioptimalkan untuk kasus penggunaan spesifik seperti *Business Intelligence* (BI) melalui Power BI, pelaporan tingkat lanjut, dan sebagai input fitur yang handal untuk model *Machine Learning Random Forest.*
b. **Proses (Transformasi Lanjutan via Apache Spark dan HiveQL):**
    1. Data dari Lapisan Silver akan diagregasi lebih lanjut dan ditransformasi untuk kebutuhan spesifik.
    2. **Agregasi Bisnis:** Kalkulasi total produksi, rata-rata produksi, tren pertumbuhan, atau metrik lain berdasarkan dimensi waktu, lokasi, dan komoditas.
    3. **Pembentukan Fitur (*Feature Engineering*):** Pembuatan fitur-fitur baru yang relevan untuk model prediksi Random Forest (misalnya, data historis produksi sebagai fitur lag, rata-rata bergerak, dll.).
    4. **Optimasi untuk Analisis (Skema Bintang):** Data akan disusun dalam Skema Bintang yang matang untuk performa *query* yang optimal dan kemudahan penggunaan di Power BI.
        a. **Tabel Sentral (Fact_Produksi):** Berfungsi sebagai repositori metrik-metrik esensial yang siap dianalisis, meliputi Luas_Panen_Aktual, Jumlah_Produksi_Aktual, dan nantinya akan diisi dengan Jumlah_Produksi_Prediksi setelah model dijalankan. Atribut satuan untuk setiap metrik juga disertakan.
        b. **Tabel Pendukung (Dim_Waktu, Dim_Lokasi, Dim_Komoditas):**
            1. Dim_Waktu: Menyajikan konteks temporal (Tahun, Kuartal, Bulan) yang final dan terverifikasi.
            2. Dim_Lokasi: Menyajikan konteks spasial (Provinsi, Kabupaten/Kota di Pulau Sumatera) yang final.
            3. Dim_Komoditas: Menyajikan konteks jenis hasil pertanian (Padi, Beras) yang final.
    5. **Dataset Khusus untuk Machine Learning:** Subset data atau fitur yang telah di-*engineer* dari lapisan Gold akan disiapkan dan diekspor dalam format yang sesuai untuk pelatihan dan evaluasi model di Spark MLlib.
c. **Penyimpanan:** HDFS (data untuk tabel Hive dalam Skema Bintang) dan potensi ekspor dataset ke HDFS dalam format lain (misalnya, Parquet atau CSV) untuk ML.

d. **Tools:** Apache Spark (untuk agregasi dan feature engineering), Apache Hive (untuk Skema Bintang), Spark MLlib (untuk konsumsi data model).

3.2. Akuisisi dan Konsolidasi Data

a. **Asal Data:** Kumpulan data mengenai produksi padi dan beras, yang mencakup informasi luas panen di seluruh **Pulau Sumatera** untuk rentang waktu 2018–2024, diperoleh dari berkas CSV yang telah disediakan. Sumber ini diasumsikan berasal dari publikasi resmi Badan Pusat Statistik (BPS) atau entitas penyedia data pertanian kredibel lainnya.
b. **Penyatuan Format:** Berkas data berformat CSV akan menjadi landasan bagi pembentukan dataset terpadu.
c. **Harmonisasi Data:** Proses penyeragaman skema data akan dilaksanakan, mencakup standarisasi penamaan kolom (seperti `Provinsi`, `Kabupaten/Kota`, `Luas Panen [Tahun]`, `Produksi Padi [Tahun]`), penetapan tipe data yang konsisten, dan penjaminan integritas format guna menghasilkan himpunan data yang komprehensif dan sahih.

3.3. Tahap Pra-Pengolahan Data (Proses ETL)

a. **Ekstraksi Informasi:** Data diekstrak dari berkas CSV sumber.
b. **Transformasi Data:**
    1. **Restrukturisasi Data (Unpivoting):** Kolom-kolom data yang merepresentasikan periode tahunan dalam CSV (contoh: `Luas Panen 2018`, `Produksi Padi 2018`, dan seterusnya) akan direstrukturisasi menjadi format baris individual. Setiap baris hasil transformasi ini akan memuat atribut `Tahun`, `Luas_Panen_Aktual`, serta `Jumlah_Produksi_Aktual` (khusus untuk komoditas Padi).
    2. **Penyaringan dan Pemurnian Data:** Apache Spark akan dimanfaatkan untuk membersihkan data dari entri yang tidak lengkap (dengan aplikasi strategi imputasi yang sesuai dan memungkinkan), menangani data yang menyimpang (anomali atau pencilan), serta mengeliminasi rekaman data yang redundan.
    3. **Identifikasi** `Komoditas_ID`**:** Untuk data yang berasal dari CSV (merujuk pada Padi), atribut `Komoditas_ID` akan disesuaikan dengan kode identifikasi untuk "Padi" dalam `Dim_Komoditas`. Data untuk "Beras" akan memerlukan alur pemrosesan tersendiri atau didasarkan pada asumsi konversi dari Padi ke Beras, jika data Beras tidak tersedia secara eksplisit.
    4. **Pencocokan dengan Tabel Dimensi (Lookup):** Atribut `Tahun`, `Provinsi`, `Kabupaten/Kota`, dan nama komoditas akan dicocokkan untuk memperoleh `Waktu_ID`, `Lokasi_ID`, dan `Komoditas_ID` yang relevan dari tabel-tabel dimensi.
    5. **Penetapan Unit Ukur:** Kolom `Satuan_Luas_Panen` (misalnya, "Hektar") dan `Satuan_Produksi` (misalnya, "Ton") akan ditambahkan berdasarkan metadata atau standar unit yang berlaku.
c. **Pemuatan Data (Load):** Data yang telah melalui proses transformasi akan dimuat ke dalam tabel `Fact_Produksi` pada platform Hive.

3.4. Konstruksi Gudang Data Menggunakan Apache Hive

a. **Repository Data (HDFS):** Himpunan data yang telah melewati tahap pra-pengolahan awal akan disimpan dalam Hadoop Distributed File System (HDFS).
b. **Definisi Skema dan Tabel pada Hive:** Apache Hive akan digunakan untuk mendefinisikan arsitektur gudang data (meliputi tabel fakta `Fact_Produksi` dan tabel dimensi `Dim_Waktu`, `Dim_Lokasi`, `Dim_Komoditas` sesuai dengan Skema Bintang) yang berada di atas lapisan data HDFS.
c. **Ingesti Data ke Hive:** Data yang telah diproses dan ditransformasi (output dari proses ETL) akan dimasukkan ke dalam tabel-tabel yang telah didefinisikan di Hive.
d. **Penjelajahan dan Verifikasi Data dengan HiveQL:** Hive Query Language (HiveQL) akan diaplikasikan untuk melakukan eksplorasi data secara mendalam, analisis agregatif (contoh: kalkulasi total produksi dan luas panen per provinsi per tahun), serta untuk memvalidasi hasil dari tahap pra-pengolahan dan menjaga konsistensi data di dalam gudang data.

3.5. Konstruksi Model Pembelajaran Mesin (Regresi Random Forest)

a. **Seleksi Variabel Prediktor (Fitur):** Variabel-variabel yang akan digunakan untuk melatih model akan dipilih dari data yang tersedia di dalam gudang data. Ini dapat meliputi `Luas_Panen_Aktual`, data historis `Jumlah_Produksi_Aktual`, serta atribut-atribut dari `Dim_Waktu` dan `Dim_Lokasi`.
b. **Penentuan Algoritma:** Model untuk memprediksi kuantitas produksi padi dan beras akan dikembangkan menggunakan algoritma Pembelajaran Mesin jenis **Regresi Random Forest**.
c. **Platform Pengembangan:** Implementasi model akan dilakukan dengan memanfaatkan pustaka **MLlib** yang terintegrasi dalam **Apache Spark**.
d. **Proses Pelatihan Model:**
    1. Data yang relevan dari Hive akan diakses dan diproses oleh Spark.
    2. Himpunan data akan dipartisi menjadi data latih (*training set*) dan data uji (*testing set*).
    3. Model Regresi Random Forest akan dilatih menggunakan data latih. Proses penyesuaian *hyperparameter* dapat dilakukan untuk optimasi performa.
e. **Pengujian Kinerja Model:** Kualitas model prediktif akan dievaluasi menggunakan data uji dengan mengacu pada metrik-metrik berikut:
    1. **Mean Absolute Percentage Error (MAPE)**.
    2. **Root Mean Squared Error (RMSE)**.
    3. Teknik *Cross-validation* akan diterapkan selama fase pelatihan untuk menjamin keandalan dan kemampuan generalisasi model.

3.6. Penyajian Visual dan Interpretasi Keluaran Menggunakan Power BI

a. **Integrasi Sumber Data:** Hasil prediksi yang dihasilkan oleh model Pembelajaran Mesin serta data historis yang relevan (termasuk `Luas_Panen_Aktual` dan

Jumlah_Produksi_Aktual) dari gudang data (Hive) akan diekspor atau dihubungkan secara langsung ke **Microsoft Power BI**.

b. **Pengembangan Dasbor Analitik:** Dasbor interaktif akan dirancang dan dibangun menggunakan Power BI dengan tujuan untuk:
    1. Menyuguhkan visualisasi tren historis luas panen dan volume produksi.
    2. Menampilkan perbandingan antara nilai produksi aktual dengan hasil estimasi model.
    3. Menyajikan metrik-metrik evaluasi performa model.
    4. Menyediakan fungsionalitas analisis *drill-down* atau penyaringan data berdasarkan dimensi waktu, lokasi, dan komoditas.
c. **Penafsiran Hasil:** Dasbor akan difungsikan sebagai alat untuk menafsirkan hasil prediksi dan menyarikan wawasan yang berguna.

3.7. Konfigurasi Lingkungan Implementasi

Seluruh alur kerja, mulai dari instalasi Hadoop Stack hingga pengembangan model, akan dijalankan dalam suatu lingkungan komputasi yang terisolasi dan portabel, memanfaatkan Docker Desktop yang beroperasi di atas WSL2 (*Windows Subsystem for Linux 2*).

4. Dataset

Jenis dan Sumber Data

Sumber data utama yang dimanfaatkan dalam pelaksanaan Tugas Besar Analisis Big Data ini adalah himpunan data yang berasal dari Badan Pusat Statistik (BPS), sebagaimana tersaji dalam file Luas Panen Padi di Pulau Sumatera 2023 - 2025 - Sheet1.csv. Dataset ini secara komprehensif mencakup informasi mengenai aspek-aspek pertanian padi di seluruh provinsi dan kabupaten/kota yang tersebar di **Pulau Sumatera** untuk rentang waktu **2018 hingga 2024**.

Data yang dianalisis mencakup informasi tentang:

- **Luas panen padi** (dalam satuan hektar), yaitu total luas lahan pertanian yang dipanen untuk komoditas padi selama satu tahun di masing-masing provinsi.
- **Wilayah administratif**, yaitu provinsi-provinsi yang berada di wilayah Sumatera.
- **Periode waktu**, yaitu dari tahun 2018 hingga 2024.

Format dan Struktur Data

Data disajikan dalam format **CSV atau Excel** dan telah disusun secara **terstruktur berdasarkan urutan waktu (time series)** dan **secara hierarki berdasarkan provinsi dan tahun**. Struktur tabel mencakup beberapa variabel utama sebagai berikut:

Variabel	Tipe data	Deskripsi
Provinsi	Kategori	Nama provinsi di Pulau Sumatera (misalnya, Sumatera Utara, Aceh, dll.).
Kabupaten/ Kota	Kategori	Menunjukkan wilayah administratif di Sumatera berisi sebaran wilayah dan identifikasi daerah penghasil padi.
Luas Panen (Tahun)	Numerik (float)	Total luas panen padi per tahun dan per provinsi, dinyatakan dalam hektar untuk per tahunnya.
Produksi Padi (Ton)	Numerik (float)	Jumlah produksi padi dalam ton untuk tahun tertentu (2018-2024)

Rentang Waktu

Periode waktu dalam dataset mencakup tujuh tahun berturut-turut, yaitu dari tahun **2018 hingga 2024**. Rentang waktu ini dipilih karena:

- **Ketersediaan dan kelengkapan data**: Seluruh data dalam periode ini tersedia secara lengkap dari publikasi resmi BPS dan dapat dijadikan sumber yang valid dan konsisten.
- **Kestabilan administratif dan kebijakan**: Rentang waktu ini cukup representatif untuk melihat dampak kebijakan pertanian dan perubahan iklim terhadap hasil panen padi di berbagai wilayah.
- **Periode analisis tren yang ideal**: Tujuh tahun merupakan durasi yang memadai untuk mengamati pola jangka menengah dan menghindari bias musiman atau fluktuasi tahunan yang ekstrem.

Dataset mencakup seluruh Provinsi di wilayah Sumatera meliputi daerah Aceh, Sumatera Utara, Sumatera Barat, Riau, Kepulauan Riau, Jambi, Sumatera Selatan, Bengkulu, dan Lampung. Dengan luasnya wilayah yang dicakup, dataset yang digunakan cukup representatif untuk analisis yang baik di tingkat provinsi maupun pada tingkat regional secara keseluruhan. Ini memungkinkan untuk melakukan pemetaan distribusi geografi serta analisis tren regional terkait ketahanan pangan dan produksi beras di Pulau Sumatera.

Secara keseluruhan data ini berguna untuk menganalisis perkembangan produksi pangan pokok di wilayah Sumatera dari waktu ke waktu. Data disajikan dalam format tabel, dengan setiap baris merepresentasikan produksi padi untuk satu kabupaten/kota dalam satu tahun, di mana setiap baris mewakili satu daerah administratif (kabupaten/kota), dan setiap kolom

mewakili tahun tertentu. Nilai yang ditampilkan merupakan jumlah produksi dalam satuan ton. Struktur data mentah ini berbentuk tabel "lebar" (*wide format*), di mana setiap tahun untuk metrik `Luas Panen` dan `Produksi Padi` direpresentasikan sebagai kolom tersendiri. Format ini akan menjadi titik awal untuk proses ETL (*Extract, Transform, Load*), di mana data akan di-*unpivot* menjadi format "panjang" (*long format*) yang lebih kondusif untuk analisis dalam *data warehouse* dan untuk keperluan pemodelan *machine learning*.

5. Tahapan Kegiatan

Dalam upaya mewujudkan **Optimasi Prediksi Produksi Padi & Beras Sumatera via Big Data Pipeline (Random Forest & Power BI)** untuk dataset periode 2018–2024, serangkaian tahapan kegiatan yang terstruktur dan intensif telah dirancang untuk dieksekusi dalam kerangka waktu **tiga minggu**. Implementasi ini akan berfokus pada pembangunan *big data pipeline* yang tangguh dan efisien, dengan memanfaatkan kapabilitas ekosistem **Apache Hadoop (HDFS), Apache Spark (untuk ETL dan MLlib), dan Apache Hive (untuk Data Warehouse)**. Proses ini akan mengolah data mentah luas panen dan produksi padi dari seluruh Pulau Sumatera, mengubahnya menjadi wawasan prediktif yang berharga.

Rangkaian kegiatan yang dirancang tidak hanya mencakup persiapan infrastruktur dan pra-pemrosesan data, tetapi juga meliputi perancangan arsitektur *data warehouse* dengan Skema Bintang, implementasi proses ETL (*Extract, Transform, Load*) yang detail termasuk *unpivoting* data, konstruksi dan validasi *data warehouse*, pengembangan model *machine learning* Random Forest Regression secara iteratif, hingga penyajian hasil analisis melalui *dashboard* interaktif menggunakan Power BI. Luaran akhir yang diharapkan dari keseluruhan tahapan ini tidak hanya berupa laporan analisis dan visualisasi tren, tetapi juga sebuah *pipeline* data yang fungsional, *data warehouse* yang terstruktur dan siap guna, model prediktif yang telah teruji, serta *dashboard* analitik yang informatif.

Berikut adalah tabel yang merincikan alokasi kegiatan yang akan dilaksanakan dalam tiga minggu tersebut:

Minggu	Kegiatan
1	**Finalisasi Desain Arsitektur Data Warehouse (Ref: Metodologi 3.1):**

	a. Penetapan Skema Bintang (`Fact_Produksi`, `Dim_Waktu`, `Dim_Lokasi`, `Dim_Komoditas`) dan atribut kunci. **Penyiapan dan Konfigurasi Lingkungan Kerja (Ref: Metodologi 3.7):** a. Instalasi dan validasi Docker Desktop, WSL2, Apache Hadoop (HDFS), Apache Hive, dan Apache Spark. **Akuisisi dan Pemahaman Dataset (Ref: Metodologi 3.2):** a. Inspeksi dan pemahaman struktur file CSV sumber (Pulau Sumatera, 2018-2024). **Inisiasi Proses ETL dan Konstruksi Awal Data Warehouse (Ref: Metodologi 3.3 & 3.4):** a. Pengembangan awal skrip ETL menggunakan Apache Spark (fokus pada ekstraksi dan logika *unpivoting* data tahunan). b. Pembuatan skema dasar dan tabel dimensi (`Dim_Waktu`, `Dim_Lokasi`, `Dim_Komoditas`) di Apache Hive.
2	1. **Penyelesaian dan Validasi Proses ETL (Ref: Metodologi 3.3):** a. Finalisasi skrip ETL Spark (pembersihan data, *lookup* ke tabel dimensi, penentuan satuan). b. Eksekusi ETL untuk memuat data lengkap ke dalam HDFS. 2. **Pengisian dan Validasi Data Warehouse (Ref: Metodologi 3.4):** a. Memuat data yang telah ditransformasi ke dalam tabel fakta (`Fact_Produksi`) di Hive. b. Melakukan validasi data dan eksplorasi awal menggunakan HiveQL untuk memastikan integritas dan

	kualitas data warehouse. 3. **Pengembangan Awal Model Prediksi Random Forest Regression (Ref: Metodologi 3.5):** a. Pemilihan fitur dari data warehouse. b. Pembagian dataset menjadi data latih dan data uji. c. Pelatihan iterasi pertama model Random Forest Regression menggunakan Spark MLlib. 4. **Evaluasi Awal Performa Model (Ref: Metodologi 3.5):** a. Mengukur kinerja model awal pada data uji menggunakan metrik MAPE dan RMSE.
3	- **Finalisasi Model Prediktif (Ref: Metodologi 3.5):** - Analisis hasil evaluasi awal, melakukan penyesuaian minor pada model jika diperlukan dan waktu memungkinkan (*hyperparameter tuning* dasar atau iterasi fitur). - Menerapkan *cross-validation* jika feasible dalam sisa waktu. - **Pengembangan Dashboard Interaktif Menggunakan Power BI (Ref: Metodologi 3.6):** - Menghubungkan Power BI ke sumber data (Hive atau data ekspor). - Merancang dan mengimplementasikan visualisasi utama untuk tren produksi, perbandingan aktual vs. prediksi, dan metrik kunci. - **Interpretasi Hasil dan Penarikan Kesimpulan (Ref: Metodologi 3.6):** - Menganalisis wawasan dari model prediktif dan dashboard Power BI. - **Penyusunan Laporan Akhir Proyek:** - Mendokumentasikan seluruh metodologi, proses, hasil, analisis, dan kesimpulan proyek.

	- Menyiapkan materi untuk presentasi akhir.

6. Target Output Proyek

Proyek "Optimasi Prediksi Produksi Padi & Beras Sumatera via Big Data Pipeline (Random Forest & Power BI)" ini bertujuan untuk menghasilkan serangkaian luaran konkret yang menunjukkan keberhasilan implementasi solusi prediksi berbasis teknologi *big data*. Target output yang akan dihasilkan adalah sebagai berikut:

1) **Infrastruktur Big Data yang Fungsional:** Sistem Apache Hadoop (mencakup HDFS untuk penyimpanan dan YARN untuk manajemen sumber daya) serta Apache Spark berhasil diinstal, dikonfigurasi, dan berjalan dengan stabil dalam lingkungan Docker Desktop dan WSL 2. Klaster dapat diakses dan dioperasikan sesuai kebutuhan.

   Data Warehouse yang Terstruktur dan Terisi: Sebuah *data warehouse* dengan Skema Bintang berhasil dibangun menggunakan Apache Hive di atas HDFS. *Data warehouse* ini akan berisi data produksi padi dan luas panen untuk seluruh Pulau Sumatera (periode 2018-2024) yang telah melalui proses ETL (*Extract, Transform, Load*) yang komprehensif (termasuk *unpivoting*, pembersihan, dan transformasi) menggunakan Apache Spark, sehingga data bersih, terstruktur, dan siap untuk analisis.
2) **Pipeline ETL yang Operasional:** Sebuah alur kerja (pipeline) ETL yang efisien dan terdokumentasi, diimplementasikan menggunakan Apache Spark, yang mampu mengambil data mentah dari sumber (CSV), melakukan transformasi yang diperlukan, dan memuatnya ke dalam *data warehouse* di Hive.
3) **Model Prediktif Random Forest Regression yang Telah Dievaluasi:** Algoritma *Machine Learning* Random Forest Regression berhasil diimplementasikan menggunakan Spark MLlib untuk memprediksi produksi padi (dan menjadi dasar untuk prediksi beras) di Pulau Sumatera. Model ini telah dilatih, diuji, dan dievaluasi secara menyeluruh menggunakan metrik performa seperti MAPE (*Mean Absolute Percentage Error*) dan RMSE (*Root Mean Squared Error*), dengan hasil yang menunjukkan akurasi dan keandalan yang memadai.

4) **Dashboard Analitik Interaktif Power BI:** Sebuah *dashboard* interaktif yang dikembangkan menggunakan Microsoft Power BI. *Dashboard* ini akan menyajikan visualisasi data historis luas panen dan produksi, perbandingan antara data aktual dengan hasil prediksi model, tren produksi, serta metrik performa model, guna mempermudah interpretasi dan mendukung pengambilan keputusan.
5) **Laporan Akhir Proyek dan Repositori Kode:** Sebuah laporan akhir yang komprehensif, mendokumentasikan seluruh tahapan proyek mulai dari perancangan, metodologi, implementasi, hasil analisis, hingga kesimpulan dan saran. Seluruh kode program yang dikembangkan (skrip ETL, kode model *machine learning*) akan terdokumentasi dan tersedia melalui repositori GitHub.

7. Tools dan Teknologi yang Digunakan

Untuk mendukung keseluruhan proses implementasi, analisis data, dan pengembangan model dalam proyek ini, akan dimanfaatkan berbagai perangkat lunak dan teknologi yang terbagi ke dalam kategori utama berikut:

7.1. Infrastruktur dan Kontainerisasi:

- **Docker Desktop + WSL2 (Windows Subsystem for Linux 2):** Platform utama untuk kontainerisasi yang memungkinkan *deployment* dan pengelolaan lingkungan Apache Hadoop, Spark, dan Hive secara efisien, konsisten, serta portabel pada sistem operasi Windows.
- **Git & GitHub:** Sistem kontrol versi untuk manajemen kode sumber proyek, kolaborasi (jika proyek tim), dan dokumentasi perkembangan kode.

7.2. Ekosistem Big Data:

- **Apache Hadoop (versi 3.x atau terbaru):** Kerangka kerja fundamental untuk penyimpanan data terdistribusi skala besar (menggunakan HDFS - *Hadoop Distributed File System*) dan manajemen sumber daya klaster (menggunakan YARN - *Yet Another Resource Negotiator*).
- **Apache Hive (versi terbaru):** Sistem *data warehouse* yang berjalan di atas Hadoop, menyediakan antarmuka mirip SQL (HiveQL) untuk melakukan kueri, analisis data terstruktur, dan mendefinisikan skema *data warehouse* (tabel fakta dan dimensi).
- **Apache Spark (versi terbaru):** Mesin pemrosesan data terdistribusi yang bersifat umum dan cepat. Akan digunakan secara ekstensif untuk proses ETL (*Extract, Transform, Load*) data, analisis data lanjutan, dan implementasi serta pelatihan model *machine learning* melalui pustaka Spark MLlib.

7.3. Pemrograman, Analitik, dan Pembelajaran Mesin:

- **Python (versi 3.8+ atau terbaru):** Bahasa pemrograman utama yang akan digunakan, terutama untuk pengembangan skrip menggunakan PySpark (antarmuka Python untuk Spark), serta untuk tugas-tugas analisis data pendukung dan otomatisasi.

- **Spark MLlib:** Pustaka *machine learning* terintegrasi dalam Apache Spark, yang akan digunakan untuk membangun, melatih, dan mengevaluasi model prediksi Random Forest Regression.
- **SQL (melalui HiveQL dan Spark SQL):** Bahasa kueri standar untuk interaksi, manipulasi, dan analisis data yang tersimpan dalam *data warehouse* di Hive serta dalam Spark DataFrames.
- **Jupyter Notebook / Visual Studio Code (VS Code):** Lingkungan Pengembangan Terpadu (IDE) yang interaktif untuk penulisan kode (Python, PySpark, SQL), eksplorasi data, visualisasi awal, dan dokumentasi analisis.

7.4. Visualisasi Data dan Pelaporan Bisnis:

- **Microsoft Power BI:** Perangkat lunak intelijen bisnis utama yang akan digunakan untuk merancang dan mengembangkan *dashboard* interaktif. Power BI akan terkoneksi dengan sumber data dari *data warehouse* untuk menyajikan visualisasi hasil analisis, tren produksi, perbandingan data aktual dan prediksi, serta metrik performa model kepada pengguna akhir.

8. Anggota Kelompok dan Pembagian Tugas

NO	Nama Anggota	Tugas
1	Pardi Octaviando	- Metodologi dan 'Tools dan Teknologi Apache - Instalasi dan penggunaan Apache Hadoop serta integrasi dengan Hive melalui HDFS
2	Kholisaturrohmah	- Pengumpulan dan Integrasi Data (mengambil data dari sumber resmi/statistik pertanian) - Visualisasi hasil prediksi dan tren historis dengan Matplotlib & Seaborn, serta dokumentasi akhir proyek

3	Elilya Octaviani	- Target Output - Instalasi dan penggunaan Apache Hadoop serta integrasi dengan Hive melalui HDFS
4	Tria Yunanni	- Tahapan Kegiatan - Implementasi eksplorasi data dan model prediktif dengan Python & Scikit-learn di Jupyter Notebook
5	Rut Junita Sari Siburian	- Latar Belakang dan Tujuan - Instalasi dan eksplorasi Apache Hive serta penulisan dan eksekusi query HiveQL
6	Seluruh Anggota	- Validasi dan Interpretasi (analisis hasil akhir) - Pra-Pemprosesan Data

9. Referensi

[1] E. Permana Yudha, A. Rohmadi, and A. T. Setyadi, "SISTEM PREDIKSI PRODUKSI PADI DI SUMATERA MENGGUNAKAN REGRESI LINEAR," *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, vol. 8, no. 1, 2025. *(Catatan: Tahun 2025 mungkin merupakan proyeksi atau kesalahan ketik, jika ini adalah publikasi aktual, tahunnya harus sudah lewat atau saat ini).*

[2] S. Hidayat Nasution, N. Irsa Syahputri, and R. Aprilia, "PENERAPAN METODE LEAST SQUARE DALAM PREDIKSI JUMLAH PRODUKSI PADI DI KABUPATEN PADANG LAWAS," *JURNAL SAINS DAN TEKNOLOGI*, vol. 7, no. 2, pp. 128-137, 2024.

[3] D. Yana Wijaya and M. Tanzil Furqon, "Peramalan Jumlah Produksi Padi menggunakan Metode Backpropagation," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 3, pp. 1129-1137, 2022. [Online]. Available: http://j-ptiik.ub.ac.id

[4] D. Bakkara, I. S. Dumayanti, and E. Rajagukguk, "PENERAPAN HOLT-WINTERS UNTUK PREDIKSI HARGA BERAS DI SUMATERA

UTARA," *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, vol. 4, no. 2, 2024. *(Catatan: Relevansi referensi ini mungkin perlu ditinjau kembali jika fokus utama proyek adalah prediksi produksi, bukan harga, dan lingkupnya kini seluruh Sumatera, bukan hanya Sumatera Utara).*

[5] P. D. B. Perteka, I. N. Gunantara, and I. B. A. Swamardika, "Analisis Big Data UMKM Menggunakan Apache Hadoop-Hive dengan Metode Forecasting Time Series," *Majalah Ilmiah Teknologi Elektro*, vol. 23, no. 2, p. 263, 2024. *(Sebelumnya [6] pada dokumen PDF asli).*

[6] T. M. Fahrudin, A. Riyantoko, K. M. Hindrayani, G. Susrama, and M. Diyasa, "Exploratory Data Analysis pada Kasus COVID-19 di Indonesia Menggunakan HiveQL dan Hadoop Environment," in *Seminar Nasional Informatika Bela Negara (SANTIKA)*, 2020. *(Sebelumnya [7] pada dokumen PDF asli)*